Validity and reliability of metrics for translation of regional anaesthesia performance from cadavers to patients

Citation for published version:
McLeod, G, McKendrick, M, Taylor, A, Lynch, J, Ker, J, Sadler, A, Halcrow, J, McKendrick, G, Mustafa, A, Seeley, J, Raju, P & Corner, G 2019, 'Validity and reliability of metrics for translation of regional anaesthesia performance from cadavers to patients', *British Journal of Anaesthesia*, vol. 123, no. 3, pp. 368-377. https://doi.org/10.1016/j.bja.2019.04.060

Digital Object Identifier (DOI):
10.1016/j.bja.2019.04.060

Document Version:
Peer reviewed version

Published In:
British Journal of Anaesthesia

BJA
British Journal of Anaesthesia

Validity and reliability of metrics for translation of regional anaesthesia performance from cadavers to patients

Journal:	*British Journal of Anaesthesia*
Manuscript ID	BJA-2019-00163-JH025.R2
Article Type:	Laboratory Investigation
Date Submitted by the Author:	27-Mar-2019
Complete List of Authors:	Mcleod, Graeme; Ninewells Hospital and Medical School, Institute of Academic Anaesthesia; McKendrick, Mel; Heriot-Watt University, Psychology Taylor, Alasdair; Ninewells Hospital and Medical School, Institute of Academic Anaesthesia Lynch, Joanna; Ninewells Hospital, Department of Anaesthesia Ker, Jean; NES Education Scotland, Postgraduate Office Sadler, Amy; Ninewells Hospital and Medical School, Institute of Academic Anaesthesia Halcrow, Jayne; Ninewells Hospital, Department of Anaesthesia McKendrick, Gary ; Optomize Ltd, Department of Engineering Mustafa, Ayman ; Ninewells Hospital & Medical School, Anaesthesia Seeley, Jonathan; Ninewells Hospital and Medical School, Institute of Academic Anaesthesia Raju , Pavan; Ninewells Hospital and Medical School, Institute of Academic Anaesthesia Corner, George; University of Dundee, Department of Bioengineering
Keywords:	Simulation, Eye tracking, Cadaver, Metrics, Regional anaesthesia

Validity and reliability of metrics for translation of regional anaesthesia performance from cadavers to patients

Graeme McLeod
Consultant Anaesthetist, NHS Tayside
Honorary Professor of Anaesthesia
Institute of Academic Anesthesia
Ninewells Hospital & University of Dundee, UK.
g.a.mcleod@dundee.ac.uk

Mel McKendrick
CEO & Research Psychologist
Optomize Ltd, Glasgow, UK
mel@optomize.solutions
Assistant Professor
Heriot-Watt University
M.McKendrick@hw.ac.uk

Alasdair Taylor
Anaesthesia SpR
Ninewells Hospital, Dundee, UK
alasdairtaylor1@nhs.net

Joanna Lynch
Consultant Anesthetist
Ninewells Hospital, Dundee, UK
Joanna.lynch@nhs.net

Jean Ker
Emeritus Professor of Medical Education
NHS Education for Scotland
Postgraduate Medical Office
Ninewells Hospital and Medical School, Dundee, DD1 9SY
jean.ker@nes.scot.nhs.uk

Amy Sadler
Anaesthesia SpR
Ninewells Hospital, Dundee, UK
amysadler1@nhs.net

Jayne Halcrow
Anaesthesia SpR
Ninewells Hospital, Dundee, UK
jaynehalcrow@nhs.net

Gary McKendrick
Researcher
Optomize Ltd, Glasgow, UK
gary@optomize.solutions

Ayman Mustafa
Consultant Anaesthetist
Ninewells Hospital, Dundee, UK
amustafa1@nhs.net

Jonathan Seeley
Consultant Anaesthetist
Ninewells Hospital, Dundee, UK
jonathan.seeley@nhs.net

Pavan Raju
Consultant Anesthesthetist
Ninewells Hospital, Dundee, UK
pbangalore@nhs.net

George Corner
Honorary Professor of Bioengineering
University of Dundee, Dundee, UK
g.a.corner@dundee.ac.uk

Correspondence to g.a.mcleod@dundee.ac.uk

Key words

Simulation
Eye tracking
Regional anaesthesia
Metrics

Running head: Validity and reliability of performance metrics

Abstract

Background: We wish to develop metrics that quantify translation of performance from cadavers to patients. Our primary objective was to develop steps and error checklists developed from a Delphi questionnaire. Our second objective was to show that our test scores were valid and reliable.

Methods: Sixteen UK experts identified 15 steps conducive to good performance and 15 errors to be avoided during interscalene block on the soft embalmed cadaver and patients. Thereafter, 6 experts and 6 novices were trained then tested. Training consisted of: psychometric assessment; an anatomy tutorial; volunteer scanning; ultrasound guided needle insertion on a pork phantom; and on a soft embalmed cadaver. For testing, participants conducted a single interscalene block on a dedicated soft embalmed cadaver while wearing eye tracking glasses.

Results: We developed a 15 step checklist and a 15 error checklist. The internal consistency of our steps measures was 0.83 (95%CI: 0.78 – 0.89) and 0.90 (95%CI: 0.87 – 0.93) for our error measures. Experts completed more steps, mean difference 3.2 (95%CI: 1.5 - 4.8), P <0.001; had less errors, mean difference 4.9 (95%CI: 3.5 – 6.3) P < 0.001; better global ratings scores, mean difference 6.8 (95%CI: 3.6 – 10.0), P < 0.001 and more eye gaze fixations, median of differences 128 (95% CI: 0 – 288), P = 0.048. Fixation count correlated negatively with steps (r = -0.60, P = 0.04) and with errors (r = 0.64, P = 0.03).

Conclusions: Our tests were valid and reliable.

High quality regional anaesthesia, performed safely, has driven changes in patient experience and postoperative pain relief after surgery. The Royal College of Anaesthetists (RCoA) oversees training standards in the UK, and its 2010 curriculum states that after completion of higher training, trainees should be able to perform plexus and regional blocks with distant supervision[1]. However, exposure to regional anaesthesia is sporadic and many trainees complete training with limited regional anaesthesia aptitudes and skills. Our experience is that only fellows in regional anaesthesia tend to achieve the desired competencies. Many trainees lack confidence[2], find interpretation of ultrasound challenging[3], display a wide intra and inter subject variability in performance[4], and may expose patients to repeated attempts, pain and harm[5]. For example, in a study of axillary brachial plexus block, trainees committed three times more errors than experts[6]. One out of six errors were sentinel errors – events that represent a serious deviation from optimal performance, jeopardize outcome, or harm patients[5].

Our wish is to train novice anaesthetists on a simulator before patient exposure, and incorporate simulator training into routine medical activity. As yet little evidence exists to justify this approach. A recent systematic review comparing simulation versus non-simulation training for regional anaesthesia showed large variability for participants, mode of simulation and outcome[7].

Our view is that in order to show any benefits of simulation we need to develop valid and reliable metrics [8, 9] that are simple to use and specifically measure translation of skills from the cadaver to the operating room.

However, current tools such as procedure duration and cumulative sum charts (CUSUM)[4] do not measure the quality of clinical intervention, direct observation of procedural skills

(DOPS) lacks reliability for formative assessment[10], and dynamic confidence scores remain unvalidated[11]. The Global Rating Scale (GRS)[12, 13 14] is summative and descriptive. Checklists identify key steps and errors before and during nerve block [5, 12, 14, 15], but are limited to patient use, and may be generic or block specific[6]. Eye gaze tracking quantitative measure of technical performance that maps fixations and saccades (movement of the pupil from one fixation to another), and offers an insight into cognitive intention during ultrasound guided regional anaesthesia (UGRA)[16-18].

We hypothesised was that our steps and error checklist scores were valid and reliable, and were better with experts compared to novices when tested on the soft embalmed Thiel cadaver, a highly durable and reliable simulator used for regional anaesthesia training[19]. Thus, the primary objective of this study was to develop steps and error checklists. Our secondary objectives were to show: validity and internal and external reliability of our checklist scores; and validity of eye tracking metrics, global rating score and self-rating questionnaires.

Methods

Development of steps and errors

After University of Dundee non-clinical Ethics Committee approval, an SpR in Anaesthesia, with a psychology degree and research experience, interviewed 7 local UGRA experts. Each interview was tape recorded and used the Consolidated Criteria for Reporting Qualitative research (COREQ), a 32-item checklist for interviews and focus groups[20]. Interviews lasted between 20 and 40 minutes. Recordings were analysed using methodological framework analysis[21] and informed a questionnaire (Online Surveys, Bristol, UK) that sought to identify generic steps and errors critical to the performance of nerve block during both cadaver simulation training and clinical practice (Table 1).

Each question had three sub-questions:

(i) How important was the item? (answered using a 5-point categorical scale - nil, minor, moderate, major, extreme importance).

(ii) What was the likelihood of causing patient harm? (answered using a 5-point categorical scale - not very likely, not likely, equally likely, likely, very likely).

(iii) What was the potential consequence of such harm? (answered using a 4-point categorical score – negligible, minor, serious, catastrophic).

The questionnaire was refined using the Delphi method, an iterative process that looks for consensus amongst experts. We asked UK consultant anaesthetists to participate who had previously been a member of faculty at our annual regional anaesthesia mastery training course on Thiel cadavers in Dundee. All were tutors or graduates of the MSc in regional anaesthesia run from the University of East Anglia, Norwich. Responses were analysed and questions rejected if inter-rater agreement was < 0.80[6]. The remaining items formed a second, smaller questionnaire that was re-distributed to the same consultants. Items achieving ≥80% agreement were included in the final steps and error checklists.

We defined psychometric tests according to the American Association of Psychologists Standards for Educational and Psychological Testing, 2014[22]. They include tests of validity, reliability and fairness. The definitions are as follows:

- Test: A device or a procedure in which a sample of an examinee's behaviour in a specified domain is obtained and subsequently evaluated and scored using a standardised process.
- Assessment: A broader term than test, commonly referring to a process that integrates test information with information from other sources

- Validity: A unitary concept which describes the degree to which accumulated evidence and theory support the interpretation of test scores for proposed uses of tests.
- Validation: The process involving accumulation of relevant evidence to provide a sound scientific basis for the proposed test interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself.
- Construct: The concept or characteristic the test is designed to measure
- Reliability/precision: A general notion of consistency of the scores across instances of the testing procedure
- Reliability coefficient: Reliability coefficient of classical test theory, i.e. the correlation between scores on two equivalent forms of the test. Three broad categories exist: (i) alternate form: (ii) test-retest and (iii) internal consistency.
- Generalizability coefficient: Ratio of universe score variance to observed score variance. Provides separate measures of components of variance
- Item response function: A model representing the increasing proportion of correct responses to an item at increasing levels of the ability or trait being measured.
- Fairness: Responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses.

Study methods entailed rater training and testing, subject training and testing and rater assessment.

Rater video training

One month before the study, the principal investigator conducted interscalene blocks on the soft embalmed cadaver under ultrasound guidance. Thirty eight videos and ultrasound movies were recorded and stored on computer hard disc. Five raters attended a training session, viewed all recordings, and were taught to recognise all steps and errors including sentinel errors, errors that may lead to patient harm. Thereafter, raters were tested on two videos and two ultrasound recordings, and completed the 15-item steps checklist and the 15-item error checklist. Successful training was defined as 80% agreement between raters.

Subject training

We conducted the study at the Centre for Anatomy and Human Identification (CAHID). The sponsor was NHS Tayside, R&D Number 2017AN04, REC Number18/WS/0082, IRAS No 243797, ISRCTN14180589. Twelve anaesthetists (6 novice and 6 experts) voluntarily opted-in and took part in the study. Novices and experts were defined according to criteria set out by Dreyfus and Dreysus[23]. Novices saw uncomplicated actions as a series of steps, had some working knowledge of key aspects of practice, were able to achieve some steps using their own judgement, but needed guidance or supervision. In contrast experts regarded themselves as individuals who had a deep understanding of regional anaesthesia, were able to take full responsibility for their own work, routinely achieved acceptable standards, saw the full picture, and modified movement patterns to address difficulties.

At the start of the study, we collected baseline demographic characteristics, including sex, handedness, age, years in anaesthesia, grade, number of supervised and unsupervised interscalene blocks performed, (0-5, 6-10, 11-30, >30), and details of undergraduate and postgraduate training. Two independent anaesthetists acted as trainers and delivered four training sessions over two consecutive days. Experts and novices were taught in groups of three, either in the morning or afternoon, over 2 days. Participants were rotated through four training stations, each lasting 20 minutes. Training stations comprised: (i) psychometric testing; (ii) a lecture describing the anatomy and performance of interscalene nerve block using ultrasound; (iii) scanning of the interscalene nerve ventral rami on a volunteer; (iv) needle insertion practice on a pork belly with embedded tendon (WetLab-MedMeat, Warwick, UK); and (v) repetitive performance of interscalene block on a soft embalmed Thiel cadaver. The subjects received the same tutorial that is given at our Ultrasound for Novice Anaesthetics Trainees course https://dihs.dundee.ac.uk/courses/anaesthetics/ultrasound-novice-anaesthetics-trainees. Our collective experience is that novices' most difficult task is visualisation of the needle at all times. This is best achieved for novices using an in-plane technique and keeping the needle as parallel as possible to the transducer elements[24].

Each subject was informed of all the necessary steps to perform and errors essential to avoid, before and during the block. A trainer was present at all stations in order to demonstrate procedures and offer guidance similar to that provided at standard UGRA teaching courses.

The soft embalmed Thiel cadaver is soaked in a salt and acid solution for 6 months before use[25]. It is soft, flexible and durable[26] and exhibits similar elasticity and strain displacement as patients[19]. Bolus injection of embalming fluid results in perineural spread similar to patients followed by tissue relaxation and fluid dispersion. These properties have allowed our research group to investigate the fundamental mechanisms underpinning differences in performance[3], and the application of mastery training to UGRA training[18]. The cadaver is highly durable, leaves no needle tracks and withstands hundreds of injections with minimal change in anatomy[19]. The legal requirements governing cadaver use are set out in the Human Tissue (Scotland) Act 2006.

Psychometric testing used a range of validated instruments (Inquisit 5 Lab, Millisecond Software, Seattle, WA) to measure mood[27], sleepiness[28], handedness[29], wakefulness[30, 31], visuo-spatial skills[32], dexterity, sustained attention[33 34] and visual scanning skills[35] (Table 1).

Subject testing

For testing, subjects were tasked to perform an interscalene block on a second cadaver. Subjects first scanned over the airway then moved the transducer laterally to sit posterior to the right sternocleidomastoid muscle. Once, the ventral nerve roots of C5 and C6 were identified, a 21g B.Braun needle was inserted in the plane of the ultrasound elements and directed, through scalenus medius, towards the junction between C5 and C6. Once the needle tip was as close to, but not touching the ventral nerve roots, 0.25 to 0.5ml of Thiel embalming solution was injected. This volume is sufficient to observe hydrolocation. No time limit was used for testing and no instruction was given. Each subject wore SMI ETG 2w wireless eye-tracking glasses (SensoMotoric Instruments, Teltow, Berlin, Germany) that were calibrated by psychologists before use. We videoed performance using a fixed camera focused on the transducer. Ultrasound and eye tracking metrics were recorded continuously. At the end of the study, participants completed a 15-item self-rating questionnaire with an 11-point anchored scale where 0 represented the worst performance and 10 represented the best performance possible. Each question was based on the list of steps selected by the Delphi process.

Rater video analysis

All test video and ultrasound recordings from all participants were examined by raters using the validated steps and errors checklists developed from our Delphi questionnaire. Our primary end point was the summation of the number of correctly performed steps and errors. Raters also completed an 8-point global rating score (GRS) with 5 anchored categories. Parameters included: preparation; asepsis; respect for tissue; time and motion; instrument handling; flow of procedure; knowledge; and overall performance. Thus, our secondary end-points were: (i) psychometric scores; (ii) GRS; (iii) eye gaze metrics; (iv) pre-procedural scanning and procedural needling time; and (v) participant self-rating of performance.

Statistics

Normality of data was assessed by the Shapiro-Wilk test. Steps and error scores and GRS scores were regarded as dependant and total scores assessed using the paired t-test. Analysis of non-parametric continuous data such as self-report measures, cognitive tests and eye gaze results used the Mann-Whitney U test and are presented as median [interquartile range]. Significant results are presented as the median (95%CI:) of the differences using the Hodges-Lehmann estimate (Graph Pad Prism 7, La Jolla, CA.) We analysed 2x2 contingency tables using the Fisher Exact Probability Test and 4x2 contingency tables with the Freeman-Halton extension of the Fisher exact probability test. Internal reliability of our steps and errors used Cronbach's alpha. External reliability was calculated using intraclass correlation (2, 5) using a 2-way random effects for absolute agreement (SPSS, v25.0.0.1, Chicago, IL). Eye tracking data was analysed using BeGaze 3.7 software (SensoMotoric Instruments). We also used generalizability statistics[36] to model the variance of items (steps and errors), raters, and subjects. **(**gtheory package (Rstudio, Version 1.1.456 – © 2009-2018 RStudio, Inc. Boston, MA) We used the "ltm" package (Rstudio) to develop a rasch model of the probability of success or failure for specific steps and errors. The mathematical function of the item characteristic curve for binary data is:

$$p = P(X = 1) = \frac{\exp\,(\theta - \delta}{1 + \exp\,(\theta - \delta)}$$

where X is a random variable indicating success or failure on the item. For steps, X = 1 indicated successful performance and X = 0 indicated failure. Inherent ability is indicated by δ on the latent variable scale, and θ is an item-parameter, generally called the item difficulty, on the same scale. In order to analyse errors, maintain the format of the item characteristic curve, and take account of the wording of the error questionnaire, we altered code entry. Errors were highlighted as X = 0. Thus, for the error item characteristic curve probability of success should be interpreted as “success in not failing to….’” or “successful avoidance of….”. Best fit of the rasch model used bootstrap goodness-of-fit Pearson chi-squared analysis. Correlation between tests was evaluated using the Spearman correlation. Kappa for multiple raters was calculated using the Online Kappa Calculator.[37]

In all analyses, a two-tailed P-value less than 0.05 indicated statistical significance. We used the Inquisit Lab 5 platform (www.millisecond.com). (https://www.psytoolkit.org)[8] to analyse psychometric tests and *Standards for Educational and Psychological Testing*[22]

Teltow, Berlin, Germany).

Power analysis: Our experience was that a difference of at least 3 errors represented a clinically meaningful difference between experts and novices. Our pilot work suggested that the standard deviation was 3 errors. Given α=0.05 and power=0.8, effect size = 1, we recruited 12 participants in order to measure the difference between two dependant means (G*power 3.1, Dusseldorf).

Results

Checklist development

Fifty five questions were identified using methodological framework analysis and divided into eight groups (Table 2): (i) position and preparation, n = 7; (ii) pre-procedural steps, n = 9; (iii) pre-procedural scanning, n = 7; (iv) needle, n = 12; (vi) needle tip, n = 11; (vii) needle tip feedback, n = 3 and (viii) local anaesthetic injection, n = 6.

The Delphi process identified 20 steps conducive to good performance and 18 errors. Five steps and 3 errors were considered not relevant to translation of performance from cadavers to patients, giving a total of 15 steps and 15 errors for testing. The reliability or internal consistency (95%CI) of our steps checklist was 0.83 (95%CI: 0.78 – 0.89) and 0.90 (95%CI: 0.87 – 0.93) for our errors checklist.

Rater video testing

Five experts were tested on two independent videos. Experts absolute agreement was 79.8% for steps and 80.5% for errors. Reliability (Kappa) was 0.62 (95%CI: 0.31 – 0.88) and 0.61 (95%CI: 0.36 – 0.86) respectively.

Subject characteristics

Experts had more experience than novices, 14.5 [8.5 – 16.8]yr vs 3.5 [3.0 – 4.8]yr, median (95%CI) of differences, 9.5 (95%CI: 2 – 16)yr, P < 0.001. All novices had performed < 5 unsupervised blocks, and all experts had performed > 30 unsupervised blocks (P = 0.002). There was no difference in age (P = 0.13), sex (p = 0.24), hand dominance (P = 1.00) or premedical qualifications (P = 1.00). Psychological, visuo-spatial and motor tests were similar in both groups (Table 3).

Subject testing – steps and errors

Nine hundred assessments of steps and 900 assessments of errors were made by 5 raters on 12 participants (6 experts and 6 novices). The external reliability of rater assessment (intraclass correlation, (95%CI)) was 0.85 (95%CI: 0.69 – 0.94), P <0.001 for steps and 0.94 (95%CI: 0.87 – 0.98) for errors, P < 0.001.

Experts performed more steps, 8.9 (3.4) vs 5.7 (3.2), mean (95%CI) difference 3.2 (95%CI: 1.5 – 4.8), P < 0.001 and had less errors, 4.1 (3.9) vs 9.0 (3.4) mean (95%CI) difference 4.9 (95%CI: 3.5 – 6.3), P < 0.004. The variance of steps, subjects, grade and raters and their combinations are shown in Table 4. Subjects, steps and their combination contributed most to measurable variance. Rasch models of the probability of success or failure for specific steps and errors are shown in Fig 1. The easiest steps (Fig 1 A, Nos. 4, 5, 8) and errors (Fig 1 B, Nos. 1, 2, 3) were associated with preprocedural scanning and identification of the nerve. The most difficult steps (Fig 1 A, 10, 13, 15) and errors (Fig 1 B, 9, 13, 15) were associated with keeping the needle tip in view at all times and recognition of needle/nerve contact and intraneural injection.

Subject testing - global rating scores

Experts had better mean (SD) GRS scores (Table 5) than novices, 26.0 (8.8) vs 19.0 (5.4), mean (95%CI) difference 6.8 (95%CI: 3.6 – 10.0), P < 0.001. Internal consistency (Cronbach's alpha) of GRS was 0.62 (95%CI: 0.16 - 0.90), P <0.001.

Subject testing - eye tracking

Experts completed steps quicker, median (95% CI:) of differences 113.9 (25.9 – 417.8) s, P = 0.04. During the scanning phase, experts glanced less often towards the monitor, median (95% CI:) of differences 4 (95% CI: 0 - 9), P = 0.04 (Table 5).

In the needling phase, expert fixation duration was longer, median (95% CI:) of differences 525 (95% CI: 83 - 993) ms, P = 0.03, and fixation count was greater, median (95% CI:) of differences 128 (95% CI: 0 – 288), P = 0.048. Expert self-report scores were better than novices, mean (95% CI:) difference 3.9 (95%CI: 3.4 - 4.5), P< 0.001).

Subject self-report measures

Experts self-report scores were better than novices. The mean (95% CI:) difference was 3.9 (95%CI: 3.4 - 4.5), P< 0.001. All questions showed differences between experts and novices (all P< 0.001).

Metric correlation

Steps and errors were negatively correlated (r = -0.94, P < 0.001). Errors correlated with procedure duration (r = 0.70, p = 0.01) and negatively with self-report scores (r = -0.59, P = 0.045). Self report scores correlated negatively with scanning dwell time (r = -0.63, P = 0.03); and glances (r = -0.68, P = 0.02). Eye gaze fixation count during needling correlated with the number of steps (r = -0.60, P = 0.04) and the number of errors (r = 0.64, P = 0.03).

Cadaver durability

Scanning and needling images from subject 1 (expert) and subject 12 (novice) are shown in Fig. 2. No difference in anatomy is seen despite injection of embalming fluid between C5 and C6 ventral nerve roots. Subject 1 aligns the needle tip precisely whereas subject 12 inserts the needle too superficially. The needle tip is not seen.

Discussion

We developed steps and error tests that were valid and reliable. Steps and error metrics, eye gaze tracking metrics, global rating scores and participant self-reporting of performance showed differences between novices and experts. These results will enable us to use our metrics to quantify the translation of training performance on the soft embalmed Thiel cadaver to clinical performance on patients.

Strengths and weaknesses of the study

The principal strengths of our study relate to the process of checklist development, quality of rater training, and depth and range of subject assessment.

Checklist development was comprehensive. We applied COREQ[20] and methodological framework analysis[21] and identified 15 steps and 15 errors from 16 UK experts. The Delphi questionnaire was designed in order that we constructed a generic questionnaire that could be applied to translation of performance from cadaver to patients. Our results suggest that rather than recommending a particular block technique, that observation of the needle and avoidance of nerve contact are most important.

Our checklists differed from others[5, 14, 15] with respect to size and content: they were up to four-times smaller than other published checklists, and tended to focus on the details of needle nerve interaction, similar to that of Cheung et al[14]. We can account for the disparity between our checklist and those of Ahmed[5] and Sultan[15] because our checklists are applicable to all blocks and, uniquely, are transferable from soft cadavers to patients, and easy to use.

The consensus of UK experts on the technical details of needle tip position are reflected by our rasch models which show the odds of skills attainment or risk of error relative to performance. Easiest steps and errors were associated with preprocedural scanning and identification of the nerve, whereas the most difficult steps were associated with keeping the needle tip in view at all times and recognition of needle/nerve contact and intraneural injection. These results echo our teaching experience and approach, and further validate our checklists.

Differences between experts and novices were clinically relevant. The mean number of steps differed by over three and steps by almost 5, and justified our small number of participants.

Validity studies[6, 13, 15, 38] comparing the performance of experts and novices do not need large numbers because the differences between groups is large, the effect size is ~1, and data is dependant because the same test cadaver is being used. However, comparison of subject performance between studies is difficult due to variable checklist composition, number of experts used, testing, size, and statistical analysis.

Global rating differences were in accord with those recorded previously[12, 14] and our self-rating questionnaire results were in keeping with this difference. Psychomotor skills did not differ between experts and novices, in contrast to the work of Shafqat et al[39] who demonstrated a negative correlation between mental rotation tests and novice needle performance on a turkey phantom. However, our sample size was five-fold smaller and our range of values was large. We will require larger samples in order to judge whether psychomotor tests impact on either scanning or needling skills during UGRA on the soft embalmed cadaver.

Our objective eye tracking results agree with our previous work on the soft embalmed cadaver that showed that expertise was associated with more focussed attention[11]. We showed that reductions in pre-procedural glances towards the screen and procedural eye fixation time were consistent with expert performance, and that eye tracking metrics discriminated between different skill levels during the scanning and needling phases of interscalene nerve block. It is likely that a lower fixation count is driven in part by reduced switches between the screen and tools and our results reflect patterns shown in laparoscopy[36-38].

Our findings show validity of eye gaze tracking: fixations and glances differentiated between experts and novices and correlated with steps and errors. Longer fixation durations evident in the expert group have been associated with more local processing and higher memory load[39]. This was more evident in the needling phase where focus was concentrated around the needle as the trajectory of the needle was followed. The lack of inherent group differences in cognitive tests indicate that steps, error and eye movement data are due to skill acquisition. Only 3 studies have incorporate eye tracking into assessment of regional anaesthesia performance. Two UGRA eye tracking studies[17, 40] recruited 6 trainees and 6 experts. The first[17] demonstrated qualitative eye gaze "heat-maps" when injecting into a gel phantom, while the second quantified trainees' interpretation of five ultrasound images[40]. A third, more extensive study, created individual learning curves for eye gaze fixation and eye glance, albeit with a wide intra-and inter-subject variation in the rate of skill acquisition, and discriminated trainee performance based on the slope and variance of the learning curve[18].

Wider impact of research

Our study is the first to validate steps and errors for the measurement of translational performance from cadavers to patients, validates eye tracking technology as a quantitative measure of regional anaesthesia performance, and informs the debate surrounding the best measure of performance reliability[41].

Our metrics will enable us to measure the translation of performance from simulators to patients and ask important questions about the impact of simulation in regional anaesthesia on clinical performance. This study is the first of three sequential studies. We will use the results of this study to power two RCTs that will (i) endeavour to compare high fidelity cadaver training vs low fidelity pork specimen training, and (ii) compare mastery vs standard teaching. The latter two studies will train subjects on cadavers and test subjects' performance on patients. Valid objective metrics will allow accurate measurement of performance to a predefined benchmark at basic, intermediate, higher and advanced levels, and offer the opportunity to standardise performance globally. Limited training resource may be more efficiently used to target areas of weakness or to optimise performance in a simulated setting prior to clinical practice.

Widespread application of eye tracking to UGRA is presently limited by the cost of equipment, and the need for an experienced psychologist to input data by hand and interpret results. Automation of eye tracking data analysis is being developed by our research group. Our intention is to enhance training experience with real time objective performance feedback in order to accelerate learning and answer key questions in UGRA robustly. The cost of eye tracking technology is falling rapidly making this affordable to an increasing number of educational institutions worldwide.

We used generalizability (G) theory[36] because it calculates the variance associated with all factors impacting on performance. It is the current recommended statistical approach used in medical educational studies. Our work showed much variance between subjects, items and their combination in both steps and error models. For comparison purposes, we purposefully measured inter-rater agreement during rater training, and calculated kappa and type 2 intraclass correlation for subject testing as recommended by others[5, 12]. We obtained good overall agreement ~ 80% during rater training, and during subject testing, kappa > 0.60 and intraclass correlation, 0.85 to 0.94. These results hide the variance attributable to subjects and checklist items as well as a large unaccountable error variance. Our approach exposes the limitations of inter-rater reliability and ICC, and suggests the need for a validated objective measurement of UGRA performance such as eye tracking which is not reliant on subjective assessment.

We are not aware of any application of rasch modelling to high stakes anaesthetic and surgical training. The logistic function modelled the probability of success and probability of failure to spot errors. Thus, we present, for the first time, a graphical means of predicting item difficulty according to skill levels. The rasch model lends itself to binary measurement of surgical and anaesthetic skills assessment and we hope this approach will prove useful to trainers and trainees alike.

In conclusion, our steps and error metrics. were valid and reliable. Eye tracking, GRS scores and self-report metrics were valid and correlated with step and error metrics. We intend to use our metrics to study the effectiveness of translation of interscalene block from the cadaver simulator to the clinical setting.

References

1. RCoA. CCT in Anaesthetics - Higher Level Training (Annex D). 2 version 1.8 ed. London, UK: Royal College of Anaesthetists, 2011; August 2010 | Version 1.8.
2. McIndoe AK. Modern anaesthesia training: is it good enough? Br J Anaesth 2012; **109**: 16-20.
3. Munirama S, Zealley K, Schwab A*, et al.* Trainee anaesthetist diagnosis of intraneural injection-a study comparing B-mode ultrasound with the fusion of B-mode and elastography in the soft embalmed Thiel cadaver model. Br J Anaesth 2016; **117**: 792-800.
4. Barrington MJ, Wong DM, Slater B, Ivanusic JJ, Ovens M. Ultrasound-guided regional anesthesia: how much practice do novices require before achieving competency in ultrasound needle visualization using a cadaver model. Reg Anesth Pain Med 2012; **37**: 334-9.
5. Ahmed OM, O'Donnell BD, Gallagher AG, Shorten GD. Development of performance and error metrics for ultrasound-guided axillary brachial plexus block. Adv Med Educ Pract 2017; **8**: 257-63.
6. Ahmed OM, O'Donnell BD, Gallagher AG, Breslin DS, Nix CM, Shorten GD. Construct validity of a novel assessment tool for ultrasound-guided axillary brachial plexus block. Anaesthesia 2016; **71**: 1324-31.
7. Chen XX, Trivedi V, AlSaflan AA*, et al.* Ultrasound-Guided Regional Anesthesia Simulation Training: A Systematic Review. Reg Anesth Pain Med 2017.
8. Gallagher AG. Metric-based simulation training to proficiency in medical education:- what it is and how to do it. Ulster Med J 2012; **81**: 107-13.
9. Barsuk JH, Cohen ER, McGaghie WC, Wayne DB. Long-term retention of central venous catheter insertion skills after simulation-based mastery learning. Acad Med 2010; **85**: S9-12.
10. Chuan A, Thillainathan S, Graham PL*, et al.* Reliability of the direct observation of procedural skills assessment tool for ultrasound-guided regional anaesthesia. Anaesth Intensive Care 2016; **44**: 201-9.
11. Byrne AJ BM, McDougall SJ. . Dynamic confidence during simulated clinical tasks. . Postgraduate Medical Journal 2005; **81**: 785-8.
12. Chuan A, Graham PL, Wong DM*, et al.* Design and validation of the Regional Anaesthesia Procedural Skills Assessment Tool. Anaesthesia 2015; **70**: 1401-11.
13. Wong DM, Watson MJ, Kluger R*, et al.* Evaluation of a task-specific checklist and global rating scale for ultrasound-guided regional anesthesia. Reg Anesth Pain Med 2014; **39**: 399-408.
14. Cheung JJ, Chen EW, Darani R, McCartney CJ, Dubrowski A, Awad IT. The creation of an objective assessment tool for ultrasound-guided regional anesthesia using the Delphi method. Reg Anesth Pain Med 2012; **37**: 329-33.
15. Sultan SF, Iohom G, Saunders J, Shorten G. A clinical assessment tool for ultrasound-guided axillary brachial plexus block. Acta Anaesthesiol Scand 2012; **56**: 616-23.
16. Borg LK, Harrison TK, Kou A*, et al.* Preliminary Experience Using Eye-Tracking Technology to Differentiate Novice and Expert Image Interpretation for Ultrasound-Guided Regional Anesthesia. J Ultrasound Med 2018; **37**: 329-36.
17. Harrison TK, Kim TE, Kou A*, et al.* Feasibility of eye-tracking technology to quantify expertise in ultrasound-guided regional anesthesia. J Anesth 2016; **30**: 530-3.
18. McKendrick M, Corner G, Tafili T*, et al.* Development of a progressive clinical expertise indicator for ultrasound-guided regional anaesthesia. Br J Anaesth 2016; **116**: e938.
19. Munirama S, Eisma R, Columb M, Corner GA, McLeod GA. Physical properties and functional alignment of soft-embalmed Thiel human cadaver when used as a simulator for ultrasound-guided regional anaesthesia. Br J Anaesth 2016; **116**: 699-707.

20. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007; **19**: 349-57.
21. Richie J, Spencer L. Qualitative data analysis for applied policy research. In: Burgess Ba, ed. Analysing Qualitative Data. London, 1994; 173-94.
22. Standards for Educational and Psychological Testing American Psychological Association, & National Council on Measurement in Education, 2014.
23. Dreyfus HL, S.E. D. Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer New York: The Free Press (Macmillan), 1986.
24. Munirama S, Joy J, Columb M*, et al.* A randomised, single-blind technical study comparing the ultrasonic visibility of smooth-surfaced and textured needles in a soft embalmed cadaver model. Anaesthesia 2015; **70**: 537-42.
25. Thiel W. [Supplement to the conservation of an entire cadaver according to W. Thiel]. Ann Anat 2002; **184**: 267-9.
26. Benkhadra M, Faust A, Ladoire S*, et al.* Comparison of fresh and Thiel's embalmed cadavers according to the suitability for ultrasound-guided regional anesthesia of the cervical region. Surg Radiol Anat 2009; **31**: 531-5.
27. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995; **33**: 335-43.
28. Akerstedt T, Gillberg M. Subjective and objective sleepiness in the active individual. Int J Neurosci 1990; **52**: 29-37.
29. Veale JF. Edinburgh Handedness Inventory - Short Form: a revised version based on confirmatory factor analysis. Laterality 2014; **19**: 164-77.
30. Thomann J, Baumann CR, Landolt HP, Werth E. Psychomotor vigilance task demonstrates impaired vigilance in disorders with excessive daytime sleepiness. J Clin Sleep Med 2014; **10**: 1019-24.
31. Lee IS, Bardwell WA, Ancoli-Israel S, Dimsdale JE. Number of lapses during the psychomotor vigilance task as an objective measure of fatigue. J Clin Sleep Med 2010; **6**: 163-8.
32. Shepard RN, Metzler J. Mental rotation of three-dimensional objects. Science 1971; **171**: 701-3.
33. Robertson IH, Manly T, Andrade J, Baddeley BT, Yiend J. 'Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. Neuropsychologia 1997; **35**: 747-58.
34. Wilson KM, Finkbeiner KM, de Joux NR, Russell PN, Helton WS. Go-stimuli proportion influences response strategy in a sustained attention to response task. Exp Brain Res 2016; **234**: 2989-98.
35. Neisser U. Decision time without reaction time: experiments in visual scanning. Am J Psychology 1963; **76**: d376 - 85.
36. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach 2012; **34**: 960-92.
37. Randolph JJ. Online Kappa Calculator [Computer software]. http://justus.randolph.name/kappa. 2008.
38. Naik VN, Perlas A, Chandra DB, Chung DY, Chan VW. An assessment tool for brachial plexus regional anesthesia performance: establishing construct validity and reliability. Reg Anesth Pain Med 2007; **32**: 41-5.
39. Shafqat A, Ferguson E, Thanawala V, Bedforth NM, Hardman JG, McCahon RA. Visuospatial Ability as a Predictor of Novice Performance in Ultrasound-guided Regional Anesthesia. Anesthesiology 2015; **123**: 1188-97.

40. Borg LK, Harrison TK, Kou A, *et al.* Preliminary Experience Using Eye-Tracking Technology to Differentiate Novice and Expert Image Interpretation for Ultrasound-Guided Regional Anesthesia. J Ultrasound Med 2017.
41. Shafqat A, Rafi M, Thanawala V, Bedforth NM, Hardman JG, McCahon RA. Validity and reliability of an objective structured assessment tool for performance of ultrasound-guided regional anaesthesia. Br J Anaesth 2018; **121**: 867-75.
36. Law, B., Atkins, M.S., Kirkpatrick, A.E., and Lomax, A.J. Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. in: Proceedings of the 2004 symposium on eye tracking research & applications. ; 2004: 41–48
View in Article | Crossref | Google Scholar
37. Wilson, M.R., McGrath, J.S., Vine, S.J., Brewer, J., Defriend, D., and Masters, R.S. Perceptual impairment and psychomotor control in virtual laparoscopic surgery. Surg Endosc. 2011; 25: 2268–2274
View in Article | Crossref | PubMed | Scopus (27) | Google Scholar
38. Wilson, M.R., Vine, S.J., Bright, E., Masters, R.S., Defriend, D., and McGrath, J.S. Gaze training enhances laparoscopic technical skill acquisition and multi-stepsing performance: a randomized, controlled study. Surg Endosc. 2011; 25: 3731–3739
https://doi.org/10.1007/s00464-011-1802-2
39. Meghanathan, R.N, van Leeuwen, C. and Nikolaev, A.R. Fixation duration surpasses pupil size as a measure of memory load in free viewing. 2014; 8: 1063.

Acknowledgments

We wish to thank the NIAA for award of the BJA/RCoA project grant (WKRO-2017-0060), and the RA-UK research grant (WKRO-2018-0016) to fund this study. We also wish to thank the staff of the Centre for Anatomy and Human Identification (CAHID) at the University of Dundee.

GM, MM, JK designed the study and wrote the paper

JL developed the checklist

AT coordinated the cadaver study

GM and GC provided technical support

AS, JH, AM, JS, and PR assessed videos

Table 1 Psychometric and cognitive tests

Table 2 Checklist development. Fifty five steps (A) and errors (B) identified and categorised into seven groups. Inter-rater agreement of iterations 1 and 2 shown in fourth and fifth columns. Only items with inter-rater reliability > 0.80 indicated consensus amongst experts. Items indicated by asterisk * chosen for final steps and error checklists. NA indicates not accepted for final checklist as not appropriate for translation from cadavers to patients.

Table 3. Psychological, visuo-spatial and motor tests. No differences between groups.

Table 4. Variance calculated from generalizability theory. Checklist items (steps and errors), subjects and their interaction comprise account for most measurable variance. Subject garde associated with more variance around errors. Raters make very small contribution to variance. Residual variance is substantial for both checklists.

Table 5. Eye tracking results of experts and novices. Experts took less time, glanced less often, fixated for longer on the needle on the monitor.

Fig 1. Rasch item characteristic curves. A steps, B errors. The x-axis represents student ability. Zero represents an average student; points to the right of the zero represent a comparatively better student and to the left, a worse student. The y -axis represents the probability that the subject conducts a steps or fails to notice an error. Curves to the left suggest easier steps and errors more likely to be noticed, and vice versa. For example, identifying a nerve is relatively easy (steps 4) whereas visualising the needle tip at all times is difficult (steps10). With regard to errors, the probability of failing to alter image depth, handle the transducer and to identify the nerve epineurium on scanning is low, whereas the probability of failing to recognise needle nerve contact (error 13), spread of fluid (error 14) or enter the nerve is high (error 15). The steeper the slope, the higher the item discrimination. Slope angles indicate that errors are more discriminatory than steps.

Fig 2.

Scanning and needling phases of interscalene block on soft embalmed Thiel cadaver performed by subject no. 1 (expert) and subject no. 2 (novice). Ventral nerve roots of C5 and C6 visible. No difference in anatomical features between images. No fluid accumulation or needling tracking. The needle shaft is well aligned and tip positioned well by expert. In contrast, needle poorly aligned by novice and traversing C6 ventral nerve root.

Function	Scale/Steps	Description
Mood	Depression Anxiety Stress Scales (DASS)[23]	Three self-report scales • Depression (dysphoria, hopelessness, devaluation of life, self-deprecation, lack of interest and inertia). • Anxiety (autonomic arousal, skeletal muscle effects, situational anxiety). • Stress (difficulty relaxing, nervous arousal, and being easily upset/agitated, irritability and impatience).
Sleepiness	The Karolinska Sleep Scale[24]	9-item scale from 1 (extremely alert) to 9 (extremely sleepy).
Handedness	The Edinburgh Handedness Scale[25]	4-item scale cored in units of 50 from -100 (always left) to +100 (always right)
Wakefulness	Psychomotor Vigilance test[26, 27]	Sensitive to sleep loss. Reaction time to appearance of red stopwatch on screen.
Visuo-spatial ability	Mental Rotation Test[28]	Mentally rotate two-dimensional images to match each other.
Dexterity	Pursuit Motor steps	Follow moving dot round circle. Performance based on time on and off target.
Sustained attention	Sustained Attention to Response Steps (SART)[29, 30]	Presented with a single digit from 1 to 9 on screen in varying font sizes. Asked to press spacebar if any digit other than 3 was seen (Go trial) and to withhold the response if digit 3 was seen (inappropriate response to NoGo).
Visual scanning	Visual scanning[31]	Participants presented with a letter matrix consisting of 25 rows of 5 letters. Stepsed with finding the target letter 'K' within 10 seconds and typing in row number within 4 seconds. Responses measured on percentage of correct responses and reaction time.

A

No.	Category	Items – steps	Delphi 1	Delphi 2	Final checklist items*
1	Positioning & Preparation	How important do you consider discussing the block with trainees beforehand?	0.75		
2		How important do you consider the pre-block positioning of the cadaver joints and limbs?	1.00	1.00	*
3		How important do you consider the positioning of the ultrasound machine	0.75		
4		How important do you consider making sure the ultrasound cables are not on the floor?	0.50		
5		How important do you consider covering the transducer with a sheath?	0.75		
6		How important do you consider sterile technique?	0.63		
7		How important do you consider making the trainee sit comfortably?	0.75		
8	Pre-procedural tasks	How important do you consider stopping before you block?	0.63		
9		How important do you consider palpation of anatomical landmarks before undertaking a block?	0.25		
10		How important do you consider flushing the needle before skin insertion?	0.63		
11		How important do you consider matching needle length to the type of block?	0.63		
12		How important do you consider the application of gel to the transducer?	0.63		
13		How important do you consider the choice of transducer?	1.00	1.00	NA
14		How important do you consider altering image depth and gain on the ultrasound machine in order to optimise the image?	1.00	1.00	*
15		How important do you consider how the transducer is handled during scanning (rotation, tilt, pressure)?	1.00	1.00	*
16		How important do you consider alignment of the transducer to the screen image?	0.75		
17	Pre-procedural scanning	How important do you consider scanning and identifying the nerve?	1.00	1.00	*
18		How important do you consider scanning and identifying the nerve epineurium?	0.63		
19		How important do you consider scanning and identifying blood vessels?	1.00	1.00	NR
20		How important do you consider scanning and identifying muscles?	0.38		
21		How important do you consider scanning and identifying muscles fascial planes?	0.63		
22		How important do you consider the level of trainee focus while conducting the pre-procedural scan?	1.00	1.00	NR
23		How important do you consider scanning proximally and distally before conducting the block?	0.88	1.00	*
24	Needle	How important do you consider the choice of needle insertion site?	0.75		
25		How important do you consider trainees looking at their hands when conducting the block?	0.75		
26		How important do you consider aligning the needle to the transducer?	1.00	1.00	*
27		How important do you consider checking the needle trajectory?	0.75	1.00	NR
28		How important do you consider aligning the needle at a tangent to the nerve?	1.00	0.82	*
29		How important do you consider optimising the visibility of the target nerve?	1.00	0.94	*
30		How important do you consider keeping the transducer still?	0.63		
31		How important do you consider identifying the entire length of the needle?	0.50		
32		In order to improve visibility of the needle, how important do you consider moving the transducer rather than the needle?	0.63		
33		How important do you consider keeping the target nerve in the middle of the screen?	0.00		
34		In order to improve visibility of the needle, how important do you consider checking the orientation of the needle with the transducer at the skin surface?	0.88	0.69	
35		How important do you consider being aware of the rate of needle insertion?	0.38		
36	Needle tip	How important do you consider identifying the needle tip before advancing the needle?	1.00	1.00	*
37		How important do you consider visualising the needle tip at all times?	0.88	0.88	*
38		How important do you consider being able to adjust the position of the needle tip?	1.00	1.00	*
39		How important do you consider needle tip migration?			
40		How important do you consider quickly regaining needle tip position when tip visibility is lost?	0.88		
41		How important do you consider being able to identify the needle tip before injection?	1.00	1.00	*
42		How important do you consider needing one needle pass?	0.00		
43		How important do you consider using hydrolocation if the needle tip is not seen?	0.63		
44		How important do you consider injecting at the best possible antomical site?	0.50		
45		How important do you consider being ambidextrous?	0.13		
46		How important do you consider being able to inject as close to but not touching the epineurium?	0.50		
47	Needle tip feedback	How important do you consider feeling fascial pops?	0.38		
48		How important do you consider recognition of needle nerve contact?	0.88	0.94	*
49		How important do you consider assessment of injection pressure?	0.38		
50	Fluid injection	How important do you consider injection of a 0.5ml to 1ml hydrolocation bolus of Thiel embalming fluid in order to confirm needle tip position	0.63		
51		How important do you consider circumferential local anaesthetic spread?	0.50		
52		How important do you consider knowing how much local anaesthetic has been injected?	0.88	0.69	
53		How important do you consider knowing where local anaesthetic has spread?	0.88	1.00	*
54		How important do you consider recognition of intraneural injection?	1.00	0.94	*
55		How important do you consider trainees communicating that they are out of their depth?	1.00	1.00	NR

B

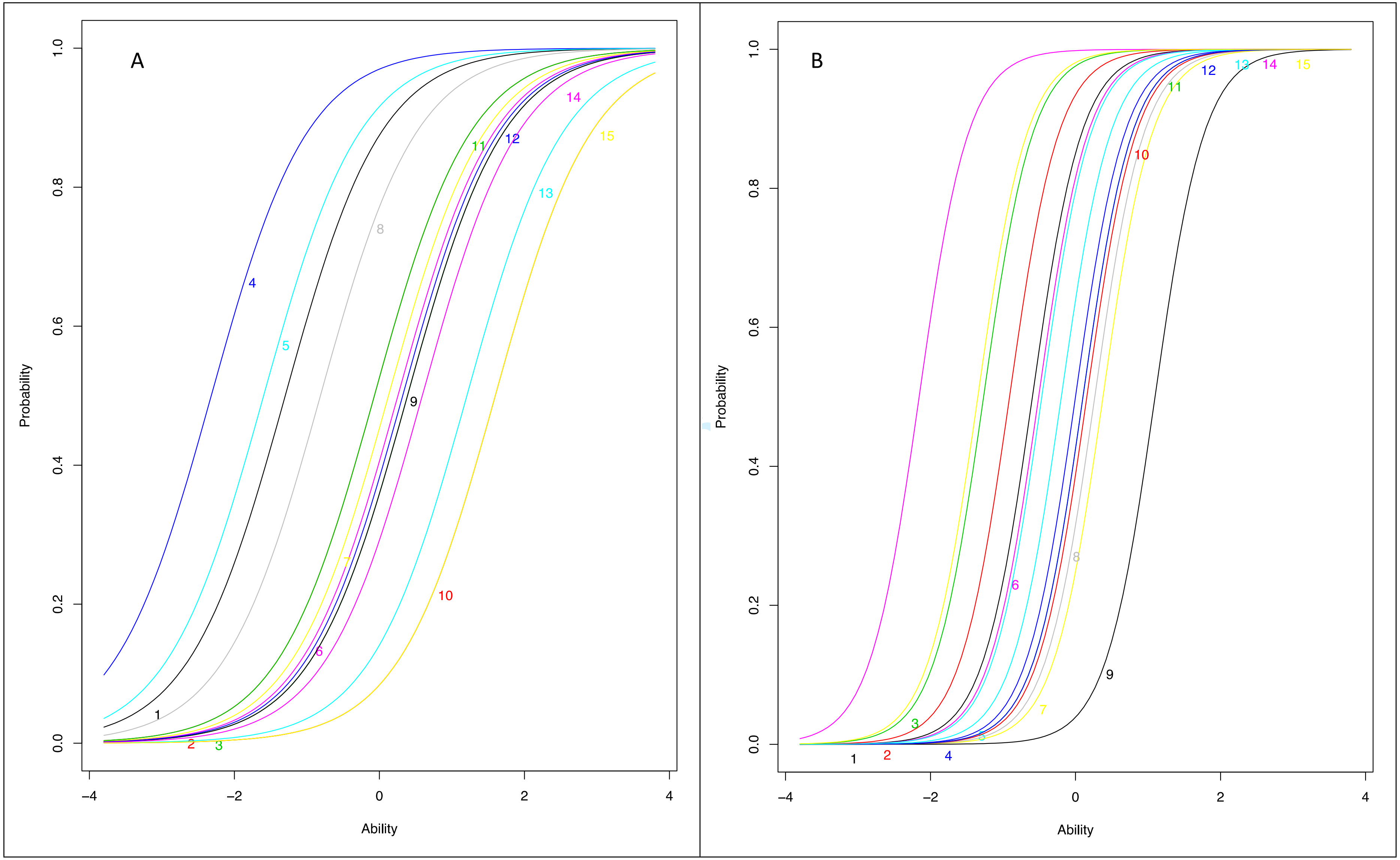
No.	Category	Items – errors including sentinel errors	Delphi 1	Delphi 2	Final checklist iitems*
	Positioning & Preparation	What would be the likelihood of causing patient harm by failing to discuss the block beforehand?	0.50	0.31	
1		What would be the potential consequence of such harm?	0.88	0.88	NA
		What would be the likelihood of causing patient harm by failing to position the joints and limbs appropriately?	0.50		
2		What would be the potential consequence of such harm?	0.75		
		What would be the likelihood of causing patient harm by failing to see the ultrasound machine and the cadaver at the same time?	0.50		
3		What would be the potential consequence of such harm?	0.75		
		What would be the likelihood of causing patient harm by failing to remove ultrasound cables on the floor?	0.13		
4		What would be the potential consequence of such harm?	0.63		
		What would be the likelihood of causing patient harm by failing to cover the transducer with a sheath?	0.50		
5		What would be the potential consequence of such harm?	0.38		
6		What would be the likelihood of causing patient harm by failing to use sterile technique?	0.50		
		What would be the potential consequence of such harm?	0.50		
7		What would be the likelihood of causing patient harm by failing to get the trainee sitting comfortably?	0.13		
		What would be the potential consequence of such harm?	0.38		
8	Pre-procedural tasks	What would be the likelihood of causing patient harm by failing to stop before you block	1.00		
		What would be the potential consequence of such harm?	0.25		
		What would be the likelihood of causing patient harm by failing to palpate anatomical landmarks before undertaking a block	0.00		
9		What would be the potential consequence of such harm?	0.00		
		What would be the likelihood of causing patient harm by failing to flush the needle?	0.25		
10		What would be the potential consequence of such harm?	0.50		
		What would be the likelihood of causing patient harm by failing to match the needle length to the type of block?	0.25		
11		What would be the potential consequence of such harm?	0.50		
		What would be the likelihood of causing patient harm by failing to apply gel to the transducer?	0.63		
12		What would be the potential consequence of such harm?	0.63		
		What would be the likelihood of causing patient harm by failing to choose the appropriate transducer?	0.38		
13		What would be the potential consequence of such harm?	0.38		
		What would be the likelihood of causing patient harm by failing to alter image depth and gain?	0.50	1.00	*
14		What would be the potential consequence of such harm?	0.88	0.94	
		What would be the likelihood of causing patient harm by failing to handle the transducer appropriately during scanning?	0.63	0.69	*
15		What would be the potential consequence of such harm?	1.00	0.94	
		What would be the likelihood of causing patient harm by failing to align the transducer to the screen image?	0.50		
16		What would be the potential consequence of such harm?	1.00		
	Pre-procedural scanning	What would be the likelihood of causing patient harm by failing to identify the nerve on scanning?	1.00		
17		What would be the potential consequence of such harm?	1.00		
		What would be the likelihood of causing patient harm by failing to identify the nerve epineurium on scanning?	0.50	0.56	*
18		What would be the potential consequence of such harm?	0.88	1.00	
		What would be the likelihood of causing patient harm by failing to identify the blood vessels on scanning?	0.88	0.81	NA
19		What would be the potential consequence of such harm?	1.00	1.00	
		What would be the likelihood of causing patient harm by failing to identify the muscles on scanning?	0.13		
20		What would be the potential consequence of such harm?	0.38		
		What would be the likelihood of causing patient harm by failing to identify the fascial planes on scanning?	0.13		
21		What would be the potential consequence of such harm?	0.38		
		What would be the likelihood of causing patient harm by failing to focus during the pre-procedural scan?	0.88	0.69	
22		What would be the potential consequence of such harm?	1.00	0.81	
		What would be the likelihood of causing patient harm by failing to scan proximally and distally before conducting the block?	0.63		
23		What would be the potential consequence of such harm?	0.63		
	Needle	What would be the likelihood of causing patient harm by failing to identify the most appropriate needle insertion site?	0.38		
24		What would be the potential consequence of such harm?	0.50		
		What would be the likelihood of causing patient harm by the trainee failing to not look at his/her hands?	0.38		
25		What would be the potential consequence of such harm?	0.63		
		What would be the likelihood of causing patient harm by failing to align the needle to the transducer?	1.00	1.00	*
26		What would be the potential consequence of such harm?	1.00	1.00	
		What would be the likelihood of causing patient harm by failing to check the needle trajectory?	0.50		
27		What would be the potential consequence of such harm?	0.63		
		What would be the likelihood of causing patient harm by failing to align the needle at a tangent to the nerve ?	0.63	0.50	*
28		What would be the potential consequence of such harm?	1.00	0.94	
		What would be the likelihood of causing patient harm by failing to optimise the visibility of the target nerve?	0.88	1.00	*
29		What would be the potential consequence of such harm?	1.00	1.00	
		What would be the likelihood of causing patient harm by failing to keep the transducer still?	0.38		
30		What would be the potential consequence of such harm?	0.63		

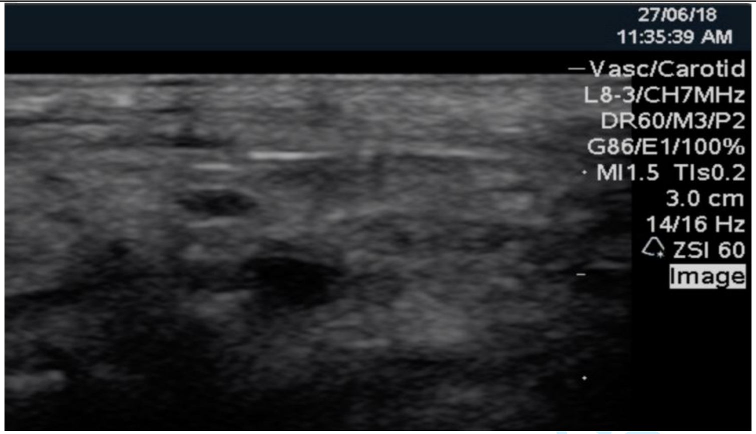
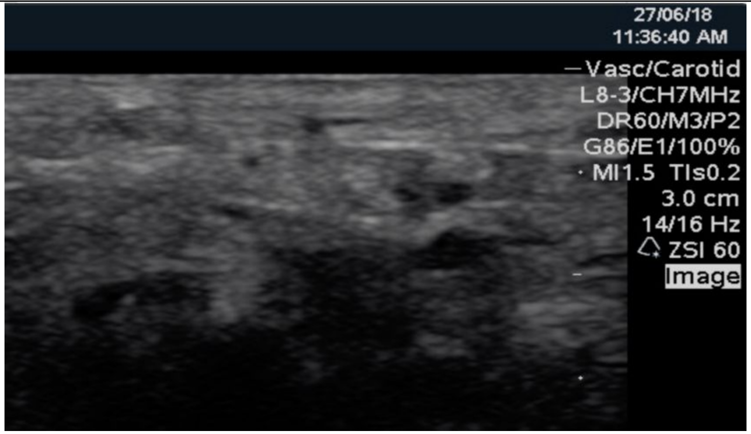
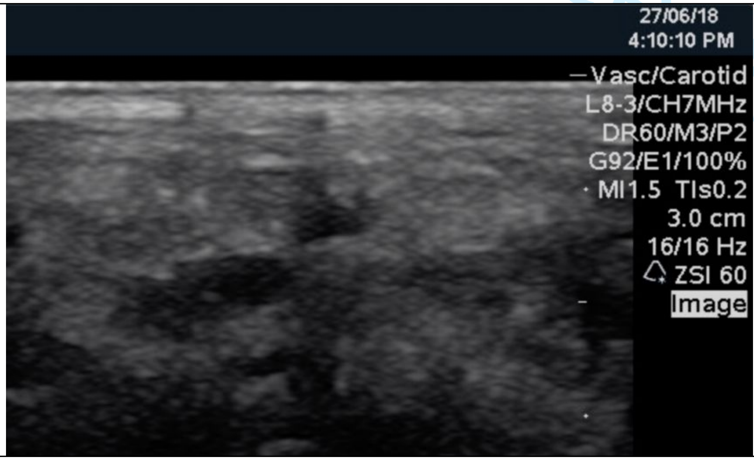
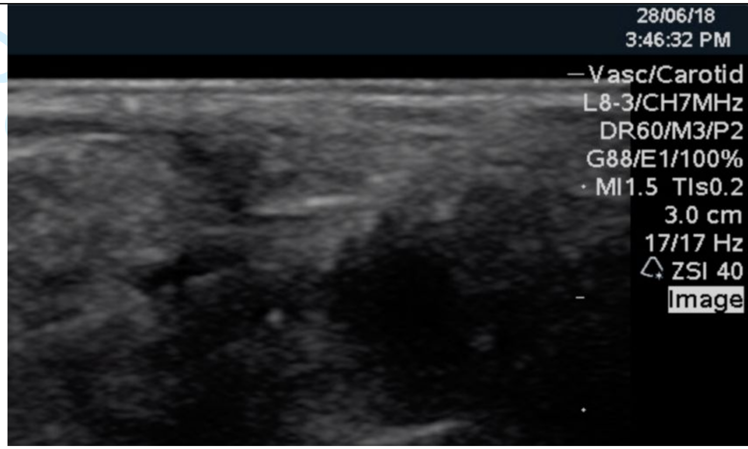
		What would be the likelihood of causing patient harm by failing to identify the entire length of the needle?	0.50	0.63	*
31		What would be the potential consequence of such harm?	0.88	1.00	
		What would be the likelihood of causing patient harm by failing to move the transducer rather than the needle?	0.50	0.38	
32		What would be the potential consequence of such harm?	0.75	0.75	
		What would be the likelihood of causing patient harm by failing to keep the target nerve in the middle of the screen?	0.13		
33		What would be the potential consequence of such harm?	0.25		
		What would be the likelihood of causing patient harm by failing to check the orientation of the needle with the transducer at the skin surface?	0.50		
34		What would be the potential consequence of such harm?	0.50		
		What would be the likelihood of causing patient harm by failing to check the rate of needle insertion?	0.38		
35		What would be the potential consequence of such harm?	0.50		
	Needle tip	What would be the likelihood of causing patient harm by failing to identify the needle tip before advancing the needle?	0.88	1.00	*
36		What would be the potential consequence of such harm?	1.00	1.00	
		What would be the likelihood of causing patient harm by failing to identify the needle tip at all times?	0.63	0.81	*
37		What would be the potential consequence of such harm?	0.88	1.00	
38		What would be the likelihood of causing patient harm by failing to identify the fascial planes?	0.50		
		What would be the potential consequence of such harm?	0.63		
39		What would be the likelihood of causing patient harm by failing to recognise needle tip migration?	1.00	0.94	*
		What would be the potential consequence of such harm?	1.00	1.00	
40		What would be the likelihood of causing patient harm by failing to quickly regaining needle tip position when tip visibility was lost?	0.88	0.82	*
		What would be the potential consequence of such harm?	0.88	1.00	
41		What would be the likelihood of causing patient harm by failing to identify the needle tip before injection?	1.00	0.94	*
		What would be the potential consequence of such harm?	1.00	1.00	
42		What would be the likelihood of causing patient harm by needing more than one needle pass?	0.00		
		What would be the potential consequence of such harm?	0.13		
43		What would be the likelihood of causing patient harm using hydrolocation if the needle tip was not seen?	0.13		
		What would be the potential consequence of such harm?	0.38		
44		What would be the likelihood of causing patient harm by failing to inject at the best possible anatomical site?	0.13		
		What would be the potential consequence of such harm?	0.13		
45		What would be the likelihood of causing patient harm by not being ambidextrous?	0.00		
		What would be the potential consequence of such harm?	0.00		
46		What would be the likelihood of causing patient harm by failing to inject as close to but not touching the epineurium?	0.00		
		What would be the potential consequence of such harm?	0.13		
47	Needle tip feedback	What would be the likelihood of causing patient harm by failing to feel fascial pops?	0.13		
		What would be the potential consequence of such harm?	0.13		
48		What would be the likelihood of causing patient harm by failing to identify needle nerve contact?	0.88	0.88	*
		What would be the potential consequence of such harm?	1.00	0.94	
49		What would be the likelihood of causing patient harm by failing to assess injection pressure?	0.50	0.38	
		What would be the potential consequence of such harm?	1.00	1.00	*
50	Fluid injection	What would be the likelihood of causing patient harm by failing to identify the hydrolocation bolus?	0.38		
		What would be the potential consequence of such harm?	0.25		
51		What would be the likelihood of causing patient harm by failing to provide circumferential local anaesthetic spread?	0.13		
		What would be the potential consequence of such harm?	0.13		
52		What would be the likelihood of causing patient harm by failing to note how much local anaesthetic had been injected?	0.50	0.63	NA
		What would be the potential consequence of such harm?	0.75	1.00	
53		What would be the likelihood of causing patient harm by failing to know where local anaesthetic has spread?	0.50		
		What would be the potential consequence of such harm?	0.63		
54		What would be the likelihood of causing patient harm by failing to identify intraneural injection?	1.00	1.00	*
		What would be the potential consequence of such harm?	1.00	1.00	
55		What would be the likelihood of causing patient harm by failing to communicate they are out of their depth?	1.00	1.00	NA
		What would be the potential consequence of such harm?	0.88	1.00	

	Novice	Expert	p
Total DASS	9 [6 - 44]	12 [2 - 36]	0.94
Depression	2 [0 - 14]	4 [0 – 28]	0.56
Anxiety	2 [0 - 6]	2 [0 - 2]	0.69
Stress	7 [2 - 26]	6 [0 - 26]	0.68
Handedness	94 [-75 - 100]	94 [-75 - 100]	1.00
Subjective sleepiness	3 [2 - 4]	2 [1 - 5]	0.18
Wakefulness	0 [0 - 1]	1 [0 - 2]	0.32
Mental rotation (% correct)	80 [70 - 90]	75 [60 - 100]	0.68
Mental rotation time (s)	4.1 [2.4 - 35.2]	8.0 [5.4 - 27.5]	0.13
Pursuit rotor. Time on target (%)	99 [96 - 100]	99 [98 - 100]	0.24
Sustained attention Time on Go Trials (s)	3.44 [3.13 – 4.83]	3.25 [2.80 – 4.28]	0.49
Sustained attention. Commission errors No-Go trials (n)	10 [4 - 15]	15 [4 - 23]	0.17
Generic visual scanning skills (%)	52 [31 - 81]	36 [0 - 64]	0.31
Generic visual scanning skills (s)	56.5 [52.0 – 69.0]	63.0 [56.9 - 75.2]	0.24

	Steps		Errors	
	Variance	% of total variance	Variance	% of total variance
Item	0.049	18.5	0.052	19.0
Subject	0.020	7.4	0.033	11.8
Grade	0.016	6.2	0.044	15.9
Rater	0.003	1.3	0.003	1.3
Subject:Item	0.017	6.3	0.030	11.0
Subject:Grade	0.010	3.8	0.012	4.3
Rater:Item	0.030	11.4	0.006	2.1
Residual	0.120	45.2	0.095	34.5

		Novice Median [Range]	Expert Median [Range]	P-value
Duration of task(s)				
Pre-Procedural		98.4 [27.3 - 282.5]	55.3 [29.2 - 87.9]	0.09
Procedural		175.4 [103.7 - 367.7]	71.3 [39.4 - 171.6]	0.04
Total		274.3 [152.9 - 650.2]	126.1 [96.4 - 232.4]	0.03
Dwell time (s)				
Pre-procedural	Monitor	70.0 [13.7 - 94.1]	48.4 [26.7 – 99.4]	0.18
	Tools	5.1 [0.1 - 80.3]	1.6 [0.5 - 8.7]	0.06
	Other	2.2 [0 - 9.7]	0.5 [0 - 2.5]	0.19
	Total	83.8 [14.5 - 160.4]	49.4 [28.7 - 73.3]	0.18
Procedural	Monitor	53.4 [17.3 - 168.3]	59.8 [36.2 - 141.1]	0.39
	Tools	11.9 [0.4 - 43.5]	9.8 [2.0 - 58.9]	0.06
	Other	0.7 [0.0 - 7.2]	0.0 [0.0 - 0.5]	0.04
	Total	69.2 [18.2 - 212.6]	85.0 [38.2 - 152.5]	0.06
Glance (n)				
Scanning	Monitor	8.5 [4 - 15]	3.5 [3 - 9]	0.04
	Tools	5.5 [1 - 9]	2.5 [1 - 6]	0.32
	Other	3.5 [0 - 11]	0.5 [0 - 3]	0.08
	Total	17 [8 - 31]	7 [5 - 18]	0.02
Needling	Monitor	12.5 [3 - 27]	4 [1 - 12]	0.06
	Tools	6 [3 - 24]	4 [1 - 8]	0.19
	Other	3 [0 - 18]	0 [0 - 8]	0.12
	Total	29 [6 - 56]	8 [2 - 28]	0.07
Fixation (ms)				
Scanning	Monitor	351.3 [194.1 - 801.8]	471.4 [401.2 - 754.9]	0.49
	Tools	215.0 [133.1 - 248.9]	223.1 [166.4 - 371.6]	0.70
	Other	192.0 [0.0 - 276.4]	115.7 [0 - 632.4]	0.95
Needling	Monitor	334.7 [187 - 923]	974.9 [710 - 1336.6]	0.03
	Tools	306.5 [217 - 522]	258.8 [133.1 - 990.0]	0.60
	Other	161.0 [0 - 182]	0 [0 - 228.4]	0.17
Fixation count (n)				
Scanning	Monitor	127 [41 - 249]	85 [38 -136]	0.18
	Tools	18 [1 - 27]	5.5 [3 - 33]	0.79
	Other	8 [0-31]	1.5 [0-10]	0.13
	Total	168 [46 - 293]	90 [44 - 179]	0.18
Needling	Monitor	160 [75-394]	60 [28-170]	0.07
	Tools	42 [16 - 131]	23 [2 - 36]	0.07
	Other	4 [0 - 35]	0 [0 - 36]	0.18
	Total	222 [91 - 530]	81 [30 - 242]	0.048


A
B
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

<table>
<tr><th></th><th>Scanning phase</th><th>Needling phase</th></tr>
<tr><td>Expert. Participant No. 1</td><td>
</td><td>
</td></tr>
<tr><td>Novice Participant No. 12</td><td>
</td><td>
</td></tr>
</table>